NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
51 HOVEY ROAD, PENSACOLA, FL 32508-1046

NAMRL Special Report 01-1

THE STATISTICAL FOUNDATIONS OF THE PILOT PREDICTION SYSTEM

D. J. Blower

ABSTRACT

The Pilot Prediction System (PPS) is a research effort designed to provide Navy policy makers with improved access to selection and training data in the aviation community. One of its main features is the ability to make predictions about the future success of aviation candidates in flight training. The purpose of this report is to present in some detail the statistical foundations of this feature of the PPS. We first describe the rudiments of statistical decision theory. Such a theory allows us to make the very best decision possible when faced with the inherent uncertainty about what is actually going to take place in the future. The second pillar on which the PPS is built is the treatment of probability from the Bayesian perspective. From readily acceptable axioms such as the sum rule and the product rule, the Bayesian approach is able to process information in a proven optimal manner. Information about an aviation candidate in the form of scores on a test battery is readily available and we would like to make the best classification of the candidate as a success or failure in flight training based on this kind of information. Specifically, the concept of a *predictive probability density* is developed within the Bayesian formalism. Once a general prediction algorithm has been constructed with the help of statistical decision theory, the details about the necessary probabilities are provided by the Bayesian approach. The prediction algorithm is one of the core modules of the PPS. It assigns candidates as likely passes or failures during some phase of flight training on the basis of preexisting data and performance on a test battery. A test battery might include such things as night visual acuity, cognitive information processing, psychomotor skills, and personality assessment. This decision to predict a pass or fail for a given candidate is taken to minimize the average monetary loss over whatever happens in the future. Two technical appendices are included for the interested reader. The first contains a simplified proof of the Bayesian predictive density that allows the prediction algorithm in the PPS to be written in its most general form. The second shows an analytical solution derived from the theory developed in the first appendix which justifies a practical approximation for predicting pass or fail during flight training for a candidate participating in a selection test battery.

Acknowledgments

I would like to acknowledge LCDR Sean Biggerstaff who first got the idea for the Pilot Prediction System off the ground, and LT Henry Williams who supported my efforts to develop a sound statistical basis for the PPS. As always, the support from the Institute for Human and Machine Cognition at the University of West Florida and its director, Dr. Ken Ford, is greatly appreciated.

INTRODUCTION

The Pilot Prediction System (PPS) is a research effort designed to provide Navy policy makers with improved access to selection and training data in the aviation community. One of its main features is the ability to make predictions about the future success of aviation candidates in flight training. Information about the goals and initial results of the PPS is referenced in Blower [1]. The purpose of this report is to present in some detail the statistical foundations behind the PPS. The theory detailed here is also applicable to *any* selection system.

We first describe the rudiments of statistical decision theory. Such a theory allows us to make the very best decision possible when faced with the inherent uncertainty about what is actually going to take place in the future. Within this theory, uncertainty about the future state of the world, or in our particular case, the future outcome of flight training for a Navy or Marine Corps Officer, is handled quantitatively. The first part of the quantitative solution chooses in favor of a prediction that results in the minimum cost. This minimum cost is formed by averaging the costs over all the possibilities that could take place. When averaging is mentioned, the notion of probability must be invoked. Uncertainty dealt with by probability theory is the second part of the quantitative solution.

The second pillar on which the PPS is built is the treatment of probability from the Bayesian perspective. From readily acceptable axioms such as the sum rule and the product rule, the Bayesian approach is able to process information in a proven optimal manner. Information about an aviation candidate in the form of scores on a test battery is readily available and we would like to make the best classification of the candidate as a success or failure in flight training based on this information. More specifically, we emphasize the concept of a *predictive probability density* as it is defined within the Bayesian formalism.

Once a general prediction algorithm has been constructed with the help of statistical decision theory, the details about the necessary probabilities are provided by the Bayesian approach. The prediction algorithm is one of the core modules of the PPS. It assigns candidates as likely passes or failures during some phase of flight training on the basis of preexisting data and performance on a test battery. A test battery might include such things as night visual acuity, cognitive information processing, psychomotor skills, and personality assessment. This decision to predict a pass or fail for a given candidate is taken to minimize the average monetary loss over whatever happens in the future.

Many numerical examples are presented the course of this paper and some care is taken to explain such concepts as a loss matrix, likelihood ratios, beta, and cut-off scores. The final example illustrates an important approximation based on the Gaussian, or Normal, curves. The case examined here is especially simple since it is based on just one composite score. A more complicated example involves the multivariate Normal when a vector of means and the covariance matrix are involved. This more difficult case is treated in Blower [1] based on the exposition by Geisser [2] and Press [3].

Two technical appendices are included at the end of the paper. Appendix A shows how the Bayesian predictive density is arrived at from simple first principles. Working out the details for this case reveals that the predictive density is merely an average likelihood for the composite score obtained by the candidate. This average likelihood is constructed from the posterior density for all the parameters that determine the likelihood. The posterior probability distribution is where the information (and uncertainty) from all the past cases is stored. In these past cases, both the data and training outcome were known. We try to leverage the information contained in the posterior density to predict a training outcome for a new candidate when, now, only the scores from the test battery are available.

The final appendix, Appendix B, presents an analytical derivation (well-known in the Bayesian literature) based on the formal theory developed in Appendix A. We show here how the Normal curves used in practice for the composite scores can be justified from the Bayesian standpoint.

STATISTICAL DECISION THEORY

Statistical Decision Theory (SDT) answers the following question: How do I take the best course of action when faced with an uncertain future? It turns out that this is the same question that selection test batteries face when trying to make the best classification of a candidate based on performance. The precepts of SDT form the fundamental basis for all selection test batteries. The PPS, as the core module for a test battery to select Navy and Marine Corps Pilots and Flight Officers, relies heavily upon SDT. The basic principle of SDT is easy to grasp.

Take that course of action which results in the minimum average loss.

A good introduction to SDT from a general statistical standpoint is contained in Berger [4]. An exposition of these principles for psychologists in the context of signal detection theory is well presented in [5]. SDT consists of three basic elements: 1) states of the world, 2) possible decisions (or actions), and 3) uncertainty about the state of the world. The standard notation for these concepts is as follows. The parameter theta, $\boldsymbol{\theta}$, stands for the set of the n states of the world,

$$\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots \theta_n\}.$$

A stands for the set of the m actions that could be taken in conjunction with the states of the world,

$$A = \{a_1, a_2 \cdots a_m\}.$$

The words *actions* and *decisions* are used synonymously in SDT. The notation $a_1, a_2, \cdots a_m$ is designed to eliminate confusion with the notation $d_1, d_2 \cdots d_m$, which is reserved for referring to data points. The uncertainty about the jth possible state of the world is captured by a probability distribution with the notation of $P(\theta_j)$.

It turns out that the main problem of concern for the PPS is especially easy for SDT. There are only two states of the world and two possible actions. Therefore,

$$\theta = \{\theta_1, \theta_2\}$$

and

$$A = \{a_1, a_2\}.$$

Here, θ_1 represents the state of the world when a flight candidate *passes* some phase of flight training, and θ_2 represents the state of the world when a flight candidate *fails* some phase of flight training. For this discussion, assume that the passing and failing criterion refers to the training that takes place through advanced flight training. Only two possible decisions will be considered in conjunction with these states of the world. They are a_1, representing the decision to *predict pass,* and a_2, representing the decision to *predict fail.*

A principle construct within SDT is the *loss matrix* for all possible combinations of θ and A. With only two states of the world and two possible actions, the loss matrix is a 2×2 matrix. Each one of the four cells of the loss matrix represents the cost in dollars for the particular combination of the state of the world and the action of that cell. The loss is expressed as $L(\theta_j, a_k)$ for the jth state of the world and the kth decision. The 2×2 matrix that the PPS uses resembles the one in Table 1. A more specifically labeled loss matrix is presented as Table 2.

Table 1: *A generic loss matrix for the decision problem concerning two states of the world and two possible actions.*

	a_1	a_2
θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Table 2: *A particular loss matrix for the decision problem concerning the selection of candidates to enter flight training. Comparing with Table 1, note that the two actions, a_1 and a_2, correspond to Predict Pass and Predict Fail, while the two states of the world, θ_1 and θ_2, correspond to Actual Pass and Actual Fail.*

	Predict Pass	Predict Fail
Actual Pass	0	C_1
Actual Fail	C_2	0

In the transition from Table 1 to Table 2, it can be seen that the two actions are *predict pass* and *predict fail* and the two states of the world are *pass advanced flight training* and *fail advanced flight training.* From the structure of the loss matrix we can observe that there are two ways to make a correct decision as well as two ways to make an incorrect decision. A correct decision occurs when 1) the action *predict pass* is taken and the state of the world is *actual pass* and 2) the action *predict fail* is taken and the state of the world is *actual fail.* For these two correct decisions a loss of 0 is assigned. Similarly, an incorrect decision occurs when 1) the action *predict fail* is taken and the state of the world is *actual pass* and 2) the action *predict pass* is taken and the state of the world is *actual fail.* For these two incorrect decisions losses of C_1 and C_2, respectively, are assigned. Subsequently, we will investigate the impact on the decisions taken when these costs are varied.

Average loss

The issue of an average loss is an important one and we address it here. In the parlance of statistics, expectation is the same as an average. For example, to average the three numbers 1,2, and 3, one adds the numbers and divides by three, yielding an average of 2. Written as a formula, this operation is

$$E(x) = \overline{x} = \frac{\sum_{i=1}^{3} x_i}{N} \tag{1}$$

where $E(x)$ is the expectation operator for x, $\overline{x}$ is the sample average, $N = 3$, and $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$.

The correct formula for the expectation of any discrete function of x is

$$E[f(x)] = \sum_{j=1}^{n} f(x_j)\, P(x_j). \tag{2}$$

The simple formula of Equation (1) has been generalized to its correct definition as a mathematical expectation of $f(x)$. In Equation (1), each x was weighted equally in finding the average. Equation (2) generalizes to the case where each function of x is weighted according to its probability of occurrence. If $f(x_j) = x_j$, $n = 3$, and $P(x_j) = 1/3$, then Equation (2) describes the same situation and gives the same answer as Equation (1).

$$E[f(x)] = (1 \times 1/3) + (2 \times 1/3) + (3 \times 1/3) = 2$$

Now that we possess a general formula for the expectation of any function, we may substitute the loss, $L(\theta_j, a_k)$, for $f(x)$ where x is now the state of the world, θ_j. By doing this, we are finding an expected loss with respect to the uncertain states of the world. If we fix k, then we are finding an expected loss for a fixed decision. There are only two decisions, so $k = 1$ or $k = 2$. Thus, we will be finding only two expected losses over the possible states of the world. The formulas for the expected loss of a predicted pass, $k = 1$, and the expected loss of a predicted fail, $k = 2$, will now be derived.

$$\text{expected loss of decision } a_k = \sum_{j=1}^{n} L(\theta_j, a_k)\, P(\theta_j). \tag{3}$$

With only $n = 2$ states of the world,

$$\begin{aligned}
\text{expected loss } a_1 &= L(\theta_1, a_1)P(\theta_1) + L(\theta_2, a_1)P(\theta_2) && (4)\\
\text{expected loss predicted pass} &= [0 \times P(\text{Pass})] + [C_2 \times P(\text{Fail})] && (5)\\
&= C_2 \times P(\text{Fail}) && (6)\\
\text{expected loss } a_2 &= L(\theta_1, a_2)P(\theta_1) + L(\theta_2, a_2)P(\theta_2) && (7)\\
\text{expected loss predicted fail} &= [C_1 \times P(\text{Pass})] + [0 \times P(\text{Fail})] && (8)\\
&= C_1 \times P(\text{Pass}) && (9)
\end{aligned}$$

We now invoke the basic principle of SDT: Choose that decision that results in the minimum expected loss. The decision rule itself is quite simple. If the expected loss of the predicted pass is less than or equal to the expected loss of the predicted fail, then predict pass, otherwise predict fail.

$$\begin{aligned}
\text{Expected loss predicted pass} &\leq \text{Expected loss predicted fail} && (10)\\
C_2 \times P(\text{Fail}) &\leq C_1 \times P(\text{Pass}) && (11)\\
\frac{C_2}{C_1} &\leq \frac{P(\text{Pass})}{P(\text{Fail})} && (12)\\
\text{If } \frac{P(\text{Pass})}{P(\text{Fail})} &\geq \frac{C_2}{C_1} && (13)\\
\text{Then} &\quad \text{Predict Pass} \\
\text{Else} &\quad \text{Predict Fail}
\end{aligned}$$

Numerical Examples

This section presents some numerical examples of the decision rule as given by Equation (13). We will first examine the case where it is more expensive to train eventual failures than to reject some candidates who would have succeeded. To solve this problem, it is first necessary to fill in the loss matrix that reflects this particular scenario. Table 3 presents an example of such a situation.

Table 3: *The loss matrix for a PPS decision problem of choosing candidates to enter flight training when it is more expensive to train eventual failures than to reject some successful candidates.*

		Predict Pass a_1	Predict Fail a_2
Actual Pass	θ_1	$L(\theta_1, a_1)$ 0	$L(\theta_1, a_2)$ \$200,000
Actual Fail	θ_2	$L(\theta_2, a_1)$ \$800,000	$L(\theta_2, a_2)$ 0

As usual, the two correct decisions do not result in any loss at all so they are assigned a cost of \$0. The incorrect decision of predicting a fail when the candidate would have passed is assessed a cost of \$200,000. This

cost is typically rather difficult to assess because it involves, among other things, the moral cost of denying a career path to a qualified candidate. However, in this example, it is viewed as a more serious incorrect decision to predict a pass when the student eventually fails training. This cost is somewhat easier to determine because it essentially involves the known training costs for a candidate through some pipeline and through some stage of training. In an ideal situation, these costs would be determined through a detailed economic analysis conducted by experts in training and selection. In any case, the exact costs are not necessary, only the ratio of the two costs. In the PPS, the main utility of the loss matrix may be in allowing users to assess the effects of different costs on the decision thresholds in a "what-if" exercise.

Given this specification of the loss matrix, we can examine the prediction algorithm of Equation (13) to see that the right hand side (rhs) of the algorithm, the ratio of the costs, has been determined. It is now necessary to assign the probabilities of pass and fail. This is where the concept of the uncertainty about the states of the world enters the picture. Whether or not a candidate is going to successfully complete training is not known beforehand. This future outcome is, by its very nature, uncertain.

A probability assignment helps capture this uncertainty. A real number (as opposed to a complex number) between 0 and 1 is assigned to the probability of a pass and then 1 minus this number is assigned to the probability of a fail. A number closer to 1 indicates more certainty about the candidate passing training, while, conversely, a number closer to 0 indicates more certainty about the candidate failing training. Numbers in the middle range close to .50 reflect a greater degree of uncertainty about the outcome.

Subjecting the candidate to a selection test battery yields relevant information that will result in a better assignment of a probability for success than if this information were not available. The PPS will incorporate all the information from all the various tests that comprise the test battery in an attempt to make an optimal assignment to the probability of a pass to any given candidate. That is, we hope that this information from the test battery will drive the probability assignment closer to 1 or 0 than if we didn't have this information. We will have much more to say on this issue in upcoming sections. For right now, we just assign these probabilities rather cavalierly in order to proceed with the numerical examples. For this first example, suppose that $P(\text{Pass}) = .75$ and $P(\text{Fail}) = .25$.

Repeating the prediction algorithm

$$\text{If } \frac{P(\text{Pass})}{P(\text{Fail})} \geq \frac{C_2}{C_1}$$

$$\text{Then} \quad \text{Predict Pass}$$

$$\text{Else} \quad \text{Predict Fail}$$

and then substituting the numbers from the above discussion results in

$$\begin{aligned} \frac{P(\text{Pass})}{P(\text{Fail})} &= \frac{.75}{.25} \\ &= 3 \\ \frac{C_2}{C_1} &= 4 \\ 3 &\geq 4 \text{ is FALSE} \end{aligned}$$

Therefore, the prediction algorithm outputs a PREDICT FAIL. In this case, a *probability* of passing equal to 75% is simply not high enough to commit to a decision to *predict* a pass. This is due to the high cost of the wrong decision to let a candidate into training when he or she fails.

Returning to first principles to calculate the expected loss for both decisions can provide a check on the correctness of the prediction algorithm, Using the abbreviation EL for expected loss, and because predicting a pass

is action a_1, we have for the $n = 2$ states of the world

$$\begin{aligned} EL(a_1) &= \sum_{j=1}^{n} L(\theta_j, a_1)P(\theta_j) \\ &= L(\theta_1, a_1)P(\theta_1) + L(\theta_2, a_1)P(\theta_2) \\ &= [\ \$0 \times P(\text{Pass})\ ] + [\ \$800,000 \times P(\text{Fail})\ ] \\ &= [\ \$0 \times .75\ ] + [\ \$800,000 \times .25\ ] \\ &= \$200,000. \end{aligned}$$

The expected loss for predicting fail, action a_2, is likewise,

$$\begin{aligned} EL(a_2) &= \sum_{j=1}^{n} L(\theta_j, a_2)P(\theta_j) \\ &= L(\theta_1, a_2)P(\theta_1) + L(\theta_2, a_2)P(\theta_2) \\ &= [\ \$200,000 \times P(\text{Pass})\ ] + [\ \$0 \times P(\text{Fail})\ ] \\ &= [\ \$200,000 \times .75\ ] + [\ \$0 \times .25\ ] \\ &= \$150,000 \end{aligned}$$

SDT says to take that decision which results in the minimum expected loss. Because the expected loss for action a_2, predict fail, is less than the expected loss for action a_1, predict pass, ($150,000 compared to $200,000) the PPS algorithm predicts fail.

If a candidate's probability of passing based on the test scores were raised to 90% from the 75% of the previous example, then

$$\begin{aligned} \frac{P(\text{Pass})}{P(\text{Fail})} &\geq \frac{C_2}{C_1} \\ 9 &\geq 4 \text{ is TRUE} \end{aligned}$$

and the candidate would be admitted into training as a predicted pass. Given the particular losses assigned to the incorrect decisions, we can see that the threshold probability of passing based on test battery data must be at least 80%. Anything lower results in the PPS predicting fail.

As the ratio of losses for these incorrect decisions climbs, it becomes even harder for a candidate to get accepted into training. As the following numerical example illustrates, an increased ratio of C_2 to C_1 raises the threshold for acceptance into training even higher. Table 4 presents a scenario when training is very expensive, but the pool of qualified applicants wishing to be trained is large. In this case, the monetary loss associated with the incorrect decision to predict a pass when the outcome is a failure in training has been raised to $1,350,000. The monetary loss with the other incorrect decision to predict a failure when the candidate would have successfully completed training is lowered to $150,000. For this altered situation,

$$\frac{P(\text{Pass})}{P(\text{Fail})} \geq 9$$

Table 4: *The loss matrix for the decision problem of choosing candidates to enter flight training when training is very expensive, but the pool of qualified applicants is large.*

		Predict Pass a_1	Predict Fail a_2
Actual Pass	θ_1	$L(\theta_1, a_1)$ 0	$L(\theta_1, a_2)$ \$150,000
Actual Fail	θ_2	$L(\theta_2, a_1)$ \$1,350,000	$L(\theta_2, a_2)$ 0

therefore, the probability of passing given performance on the test battery must be 90% or greater for the decision rule to recommend acceptance into training. Going back to first principles, this means that the average loss for predicting a pass for any probability of passing less than 90% is greater than the average loss of predicting a fail.

Of course, the change in the decision criterion can work just as well in the opposite direction. If the losses for the two incorrect decisions were judged to be of equal value as in Table 5 below, then the ratio of costs would change to

$$\frac{C_2}{C_1} = 1.$$

As soon as the ratio of the probability of passing to the probability of failing,

$$\frac{P(\text{Pass})}{P(\text{Fail})}$$

becomes greater than 1, a predicted pass results. A probability of passing equal to 50% or greater would be sufficient to predict a pass in this case.

Table 5: *The loss matrix for the decision problem of choosing candidates to enter flight training when it is equally as costly to reject someone who could have passed as it is to admit someone who fails.*

		Predict Pass a_1	Predict Fail a_2
Actual Pass	θ_1	$L(\theta_1, a_1)$ 0	$L(\theta_1, a_2)$ \$500,000
Actual Fail	θ_2	$L(\theta_2, a_1)$ \$500,000	$L(\theta_2, a_2)$ 0

The threshold used by the PPS for admitting a candidate moves in the expected direction. When training costs are high and the pool of applicants is large so that the replacement cost for rejected candidates is lowered, better scores on the test battery are demanded. The threshold is placed at a higher level; for example, the probability of passing based on the test battery results might have to be higher than 90%. On the other hand, if for whatever reason the cost of rejecting someone who later would have passed is comparable to the training costs, then the threshold is lowered and the decision to accept a candidate could be lowered to a probability of passing based on test battery results of 50%.

THE BAYESIAN APPROACH

We now address the second pillar of the statistical foundations of the PPS. Up to this point, we have used a loose notation for $P(\text{Pass})$ and $P(\text{Fail})$. Each of these probabilities is actually conditioned on data from the test battery, so the notation must express this fact.

During the development of the PPS (or any selection test battery), there is a validation phase where data are collected on N subjects. No actual selection takes place, all N candidates participate in the test battery and then enter training whatever their score. At some point, the validation phase is completed and the system is then brought into an operational mode where actual selections take place. Assuming this occurs for the $(N+1)$st candidate, the correct notation for the probability of pass and probability of fail in the prediction algorithm should be

$$P(\text{Pass}|D_{N+1}, D_N) \text{ and } P(\text{Fail}|D_{N+1}, D_N).$$

Here $D_1, D_2, \cdots D_N$ is the notation for the data in the database from the N candidates for whom both the scores on the test battery and the training outcome are known. Then, D_{N+1} is the notation for the data of the candidate currently being tested and whose training outcome is unknown.

The addition of the solid vertical line to a general probability expression like $P(A|\mathcal{I})$ is the "conditioned upon" symbol. It says that the assigned probability for the proposition that appears to the left of the symbol, A, is conditioned on the truth of what appears to the right of the symbol, $\mathcal{I}$. For our current application, this means that any probability for passing or failing is conditioned upon both the receipt of the $(N+1)$st candidate's data from the test battery and what has happened to the previous N candidates. It indicates that this probability of passing or failing may be, and most likely is, different from any other probability conditioned on something else. For us, the most relevant comparison is when the probability assignment conditioned on the test battery data is contrasted with a probability assignment not based on the selection system. In other words, the question is whether or not the selection system adds relevant information that raises or lowers the probability of passing.

The Bayesian approach is specifically designed to handle this situation. It uses the basic axioms from probability theory to find an updated probability for the $(N+1)$st candidate conditioned on the test scores achieved by this candidate. One easy result from the application of the basic axioms of probability theory is Bayes's Theorem, written as

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{14}$$

The proposition H usually consists of a number of mutually exclusive and exhaustive hypotheses, labeled $H_1, H_2, \cdots H_j, \cdots H_K$. Bayes's Theorem in this form is used to find the probability of the ith hypothesis, H_i, as conditioned on the data, D. Then Bayes's Theorem is rewritten as

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum_{j=1}^{K} P(D|H_j)P(H_j)} \tag{15}$$

where the denominator has been expanded as the sum of all the terms that could appear in the numerator.

Our current application presents us with only two mutually exclusive and exhaustive hypotheses

$$H_1 = \text{ PASS and } H_2 = \text{ FAIL}$$

and, D, the data, is actually the data from the $(N+1)$st candidate, D_{N+1}. Equation (15) then assumes the forms shown below in Equations (16) and (17) for each of the two hypotheses,

$$P(\text{Pass}|D_{N+1}) = \frac{P(D_{N+1}|\text{Pass})\ P(\text{Pass})}{P(D_{N+1}|\text{Pass})\ P(\text{Pass}) + P(D_{N+1}|\text{Fail})\ P(\text{Fail})} \tag{16}$$

$$P(\text{Fail}|D_{N+1}) = \frac{P(D_{N+1}|\text{Fail})\ P(\text{Fail})}{P(D_{N+1}|\text{Pass})\ P(\text{Pass}) + P(D_{N+1}|\text{Fail})\ P(\text{Fail})} \tag{17}$$

We temporarily suppress explicitly writing out the information from the previous data, D_N, to keep the formulas manageable. Later, we will reintroduce the symbol to see how the Bayesian formalism handles D_N.

The prediction algorithm up to now has been written as

$$\text{If } \frac{P(\text{Pass})}{P(\text{Fail})} \geq \frac{C_2}{C_1}$$

Then Predict Pass

Else Predict Fail

The next step is to substitute for $P(\text{Pass})$ and $P(\text{Fail})$ the more correct probabilities conditioned on the data obtained by the candidate that is to be classified. The goal is to form the ratio

$$\frac{P(\text{Pass}|D_{N+1})}{P(\text{Fail}|D_{N+1})}.$$

From Equations (16) and (17) the respective denominators cancel out, leaving

$$\frac{P(\text{Pass}|D_{N+1})}{P(\text{Fail}|D_{N+1})} = \frac{P(D_{N+1}|\text{Pass})\ P(\text{Pass})}{P(D_{N+1}|\text{Fail})\ P(\text{Fail})}. \tag{18}$$

The prediction algorithm now reads

$$\text{If } \frac{P(\text{Pass}|D_{N+1})}{P(\text{Fail}|D_{N+1})} \geq \frac{C_2}{C_1} \text{ then Predict Pass, otherwise Predict Fail.} \tag{19}$$

Many of the prediction algorithms in the literature stem from this derivation. However, the nomenclature that is employed can obscure this fact. One popular form of a prediction algorithm is phrased in terms of a *likelihood ratio* and a *response threshold.* Our prediction algorithm can be expressed in these terms, and the attempt is now made to allow the reader to translate from one symbology to the other.

Substituting Equation (18) into the revised prediction algorithm of Equation (19) yields,

$$\frac{P(D_{N+1}|\text{Pass}) \times P(\text{Pass})}{P(D_{N+1}|\text{Fail}) \times P(\text{Fail})} \geq \frac{C_2}{C_1} \tag{20}$$

$$\frac{P(D_{N+1}|\text{Pass})}{P(D_{N+1}|\text{Fail})} \geq \frac{C_2}{C_1} \times \frac{P(\text{Fail})}{P(\text{Pass})} \tag{21}$$

The left-hand side (lhs) of Equation (21) is in the form of a likelihood ratio because it is the ratio of the probability density of the data given that it came from the Pass group over the probability density of the data given that it came from the Fail group. This likelihood ratio is written as,

$$\mathcal{L}(x) \equiv \frac{P(D_{N+1}|\text{Pass})}{P(D_{N+1}|\text{Fail})}.$$

The rhs of Equation (21) is a function of the costs of making correct and incorrect decisions and the prior odds of failing over passing. Together they make up β, the response threshold, where

$$\beta \equiv \frac{C_2}{C_1} \times \frac{P(\text{Fail})}{P(\text{Pass})}.$$

Equation (21) therefore represents a decision algorithm in the form of

$$\text{If } \mathcal{L}(x) \geq \beta \text{ then predict Pass, otherwise Predict Fail} \tag{22}$$

Numerical Examples

The rhs of the new prediction algorithm, Equation (21), is dealt with first. That is, we calculate the value of β. Then we turn to the lhs of Equation (21), the likelihood ratio. Take the costs C_1 and C_2 to be in the ratio of 3:1 so that

$$\frac{C_2}{C_1} = 3.$$

For example, suppose that an attrition during advanced flight training costs \$900,000 and the cost of replacing a PPS rejected candidate is \$300,000. P(Pass) and P(Fail) are probabilities not conditioned upon any data from a selection test battery. We may assign these probabilities to be reflective of the historical record for passing or failing some phase of flight training given that no selection test battery is operational. By the phrase "no selection test battery operating," we mean any additional selection other than what is implemented by the current medical standards and intelligence tests. For example, Blower [6] has shown that psychomotor skills, scores from pre-flight ground school, and personality evaluations represent cogent information that would change the assigned probability of success. So, the assignments to P(Pass) and P(Fail) do not reflect this kind of extra information that could be used by the PPS. They are simply probabilities assigned to Pass and Fail that do not take into account scores from a test battery, for example. From the past historical records, an assignment of

$$P(\text{Pass}) = .75 \text{ and } P(\text{Fail}) = .25$$

is a fairly accurate statement through advanced flight training.

The rhs of the revised form of the prediction algorithm is then

$$\begin{aligned} \beta &= \frac{C_2}{C_1} \times \frac{P(\text{Fail})}{P(\text{Pass})} \\ &= \frac{3}{1} \times \frac{.25}{.75} \\ &= 1. \end{aligned}$$

With this particular value of β determined, the prediction algorithm looks like

If $\mathcal{L}(x) \geq 1$ then predict Pass, otherwise Predict Fail

We now turn to address the lhs of the prediction algorithm. For a new candidate, this is the ratio of the likelihood of the data from whatever additional information makes up the new selection test battery. The PPS would calculate $\mathcal{L}(x)$ as well as β for the new candidate who is to be classified. Specifically, the ratio of the likelihood of these data conditioned upon being in the PASS group must be compared to the likelihood of these data conditioned upon being in the FAIL group.

For a simple first example that seems to work well in practice, consider that these likelihoods are well described by two Normal probability density functions. The shorter word *curve* will be used as a synonym for probability density function. Even if the Normal curves are not analytically the correct curves, they are, in many cases, good approximations to the correct underlying curves. For example, this situation occurs when a composite score is constructed from many other scores in the test battery. Using a technique like *Discriminant Analysis,* such a composite score can be constructed such that the means of two curves are as far apart as possible while, at the same time, the standard deviations are made equal. These composite scores are also assumed to follow a Normal curve for both the PASS and FAIL groups. After examining the implications of the prediction algorithm for this simplification, the technical justification will be provided in Appendix B.

The discussion about the prediction algorithm is easier to follow when accompanied by sketches of these Normal curves and the placement of the response threshold. See Fig. 1 for an initial orientation. In this sketch,

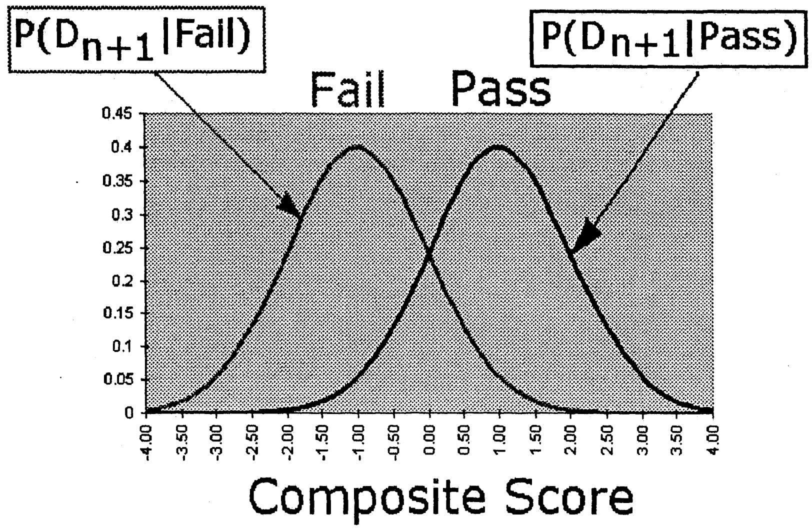

Figure 1: *Normal curves used as approximations for the likelihood of the composite score (D_{N+1}) under the PASS and FAIL assumptions.*

there is one Normal curve at the left to represent the predictive distribution of the composite score given that the candidate is a failure. Displaced to the right, is the comparable Normal curve representing the predictive distribution of the composite score assuming the truth of the candidate passing. The likelihood for each curve is simply the ordinate, or the value at the y-axis, on the Normal curve.

The candidate participates in the selection test battery and receives a specific composite score. Drawing a line upwards at the point on the x-axis representing this composite score intersects both Normal curves. For a "low" composite score, the line first intersects the Normal curve for the PASS group and then intersects the Normal curve for the FAIL group. The y-axis values at these two intersection points are the likelihoods we need to form the ratio, $L(x)$ as shown in Fig. 2. The low composite score is given a value of -1.00 in Fig. 2.

For a "low" composite score, the y-axis value is going to be larger for the FAIL group than for the PASS group. Therefore, the likelihood ratio, $L(x)$, because it is a ratio of the likelihood of the PASS group over the FAIL group, must be less than 1. As the composite score gets better and better, it eventually reaches the point where the two curves intersect. Here the y-axis values for both curves are equal and therefore $L(x) = 1$. We can imagine this process of obtaining a better composite score continuing so that eventually the line first intersects the FAIL Normal curve followed by intersecting the PASS Normal curve. Then the y-axis value for the PASS group is larger than for the FAIL group and the likelihood ratio, $L(x)$, becomes greater than 1. When a composite score is high enough to reach this likelihood ratio, then it exceeds the value of the response threshold, $\beta = 1$. At this point, the prediction algorithm will output a PREDICT PASS. If the composite score obtained by the candidate is not high enough to reach this response threshold, then the likelihood ratio is less than 1 and the prediction algorithm will output a PREDICT FAIL. See Figure 3 for an illustration of this concept.

Finding the Threshold Score

The dividing line when $L(x) = 1$ in Fig. 3 where PREDICT FAIL is distinguished from PREDICT PASS occurs at the threshold score of 0. How do you go about actually finding that particular composite score? We will answer that question shortly in a long algebraic derivation, but it is easy to see from Fig. 3 that x lies halfway between the mean of the FAIL group and the mean of the PASS group. In Fig. 3, the mean of the FAIL group occurs at -1.00 and the mean of the PASS group occurs at $+1.00$, so the threshold score is placed halfway between them at a composite score of 0.

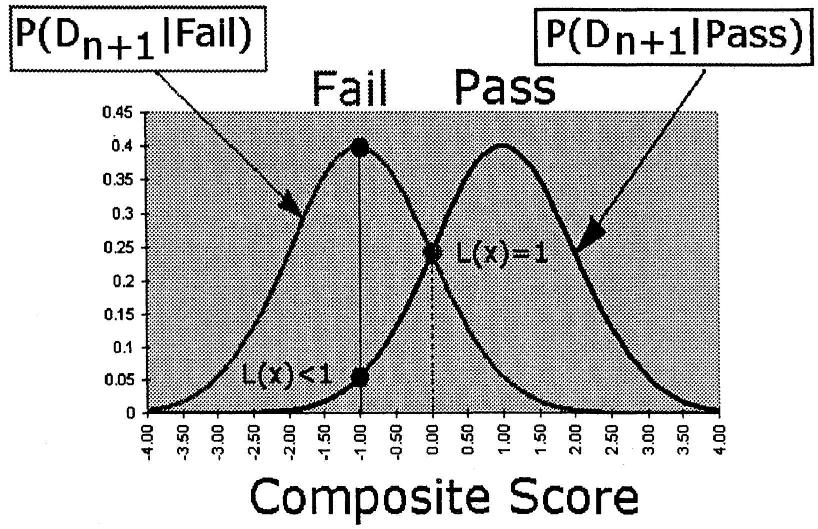

Figure 2: *Normal curves used as approximations for the likelihood of the composite score (D_{N+1}) under the PASS and FAIL assumptions. A likelihood ratio less than 1 and equal to 1 are drawn into the sketch.*

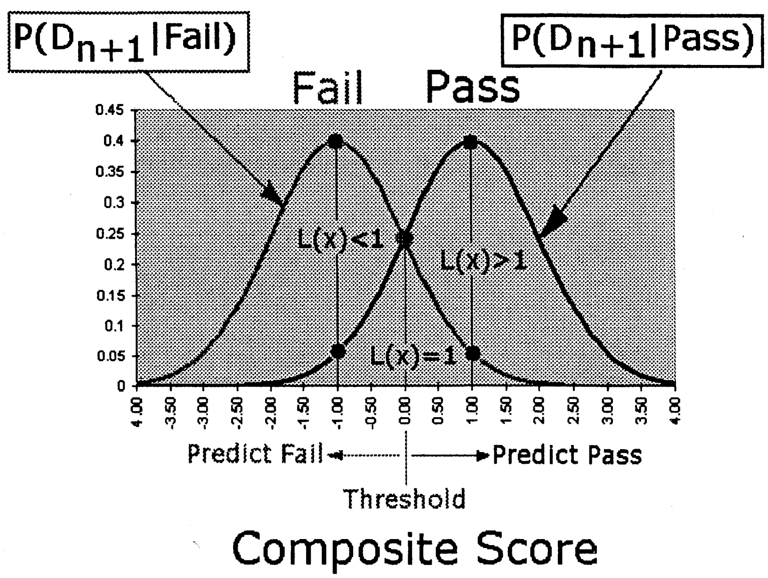

Figure 3: *Normal curves used as approximations for the likelihood of the composite score (D_{N+1}) under the assumptions of PASS and FAIL. Three likelihood ratios, the first less than 1, the second equal to 1, and the third greater than 1 are sketched in the figure. The regions for the composite score to yield a predict pass or a predict fail for the case where $L(x) = 1$ are also shown.*

We have talked about the ordinate of the Normal curve as the y-axis value. The formula for the ordinate of the Normal curve is

$$y\text{-axis value} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\} \tag{23}$$

We have already discussed the fact that the threshold response for $\beta = 1$ occurs where the two Normal curves intersect. Therefore, their ordinates must be the same value.

$$\frac{1}{\sqrt{2\pi}\sigma_F} \exp\left\{ -\frac{1}{2} \left(\frac{x-\mu_F}{\sigma_F} \right)^2 \right\} = \frac{1}{\sqrt{2\pi}\sigma_P} \exp\left\{ -\frac{1}{2} \left(\frac{x-\mu_P}{\sigma_P} \right)^2 \right\} \tag{24}$$

Discriminant Analysis returns composite scores with

$$\sigma_F = \sigma_P = 1$$

permitting the leading term to be canceled. Equation (24) then simplifies to

$$\exp\{-1/2\,(x-\mu_F)^2\} = \exp\{-1/2\,(x-\mu_P)^2\}. \tag{25}$$

Take the natural logarithmic transform of both sides of Equation (25) to yield

$$-1/2\,(x-\mu_F)^2 = -1/2\,(x-\mu_P)^2. \tag{26}$$

Now expand the quadratic on each side of the side of Equation (26)

$$-1/2\,(x^2 - 2x\mu_F + \mu_F^2) = -1/2\,(x^2 - 2x\mu_P + \mu_P^2). \tag{27}$$

Multiply the terms in the parentheses by $-1/2$

$$-1/2x^2 + x\mu_F - 1/2\mu_F^2 = -1/2x^2 + x\mu_P - 1/2\mu_P^2. \tag{28}$$

Cancel the leading term

$$x\mu_F - 1/2\mu_F^2 = x\mu_P - 1/2\mu_P^2. \tag{29}$$

Collect like terms on the appropriate sides of Equation (29)

$$x(\mu_P - \mu_F) = 1/2\,(\mu_P^2 - \mu_F^2). \tag{30}$$

The rhs of Equation (30) can be decomposed into

$$(\mu_P^2 - \mu_F^2) = (\mu_P - \mu_F)(\mu_P + \mu_F) \tag{31}$$

leading to

$$x\,(\mu_P - \mu_F) = 1/2\,(\mu_P - \mu_F)(\mu_P + \mu_F). \tag{32}$$

The final step is to isolate x and achieve the intuitive answer that, yes, the threshold composite score does lie midway between the means of the PASS and FAIL groups.

$$x = 1/2\,(\mu_P + \mu_F) \tag{33}$$

Changing the Response Threshold

The response threshold, β, will change as a function of different costs and different prior probabilities. β could easily change from the value of 1 as used in the numerical example given above. What effect does changing β have on the prediction as given by PPS?

In the definition of β, there are two ratios: 1) the ratio of costs, C_2/C_1; and 2) the ratio of prior probabilities, $P(\text{Fail})/P(\text{Pass})$. Again, the phrase *prior probability* is used only as a shorthand for a probability that must be assigned without having the benefit of any selection test battery scores. It will, in fact, be based on other information and should properly be written as $P(\text{Pass}|\mathcal{I})$ to indicate that a probability for passing has been assigned given the truth of some information $\mathcal{I}$.

We consider each ratio making up β in turn, and while changing one ratio, the other ratio is assumed fixed and nonchanging. As the ratio of costs increases, it becomes relatively more expensive to experience a failure during flight training than to replace a PPS rejected candidate who would have succeeded. Suppose that C_2/C_1 moves from 3 to 6 to 9 to 12, while $P(\text{Fail}|\mathcal{I})$ divided by $P(\text{Pass}|\mathcal{I})$ remains fixed at $1/3$. Then β increases from 1 to 2 to 3 to 4. The likelihood ratio, $L(x)$, must now exceed an increasingly larger number in order to predict a pass. The y-axis value of the threshold composite score for the PASS group must get larger relative to the y-axis value of the FAIL group, and this can only be accomplished if the threshold composite score is increasing. It thus becomes more difficult for a candidate to get selected because of the higher score needed for the PPS to recommend entry into training. Conversely, if the ratio of costs decreases, then, by the same reasoning, the composite score needed for the PPS to recommend entry is driven lower.

In the first case, the objective is to decrease the number of failures in training at the expense of increasing the number of incorrect rejections. In the second case, the objective is to decrease the number of candidates incorrectly rejected at the expense of increased training failures. Thus creates a continuous trade-off between these two errors as we change the cost structure in the loss matrix. Any point during this trade-off is justified by the lower expected loss of taking this action.

Due to the multiplicative nature of β, an entirely analogous event takes place if we keep the cost structure fixed and manipulate the ratio of prior probabilities. If, independently of any selection test battery scores, the probability of passing is very high, say $P(\text{Pass}|\mathcal{I}) = .90$, then the second ratio becomes $1/9$. β then moves lower in comparison with the situation in the example just discussed where $P(\text{Pass}|\mathcal{I}) = .75$ (assuming that the costs remain fixed). A lower composite score than before can result in a predicted pass. In a sense, the test battery becomes less relevant when the probability of passing without the test scores is already very high. On the other hand, when the probability of passing prior to the implementation of selection test battery is low, the threshold composite score is raised. Achieving a good score on the selection test battery then becomes more important.

REFERENCES

1. Blower, D. J. *Probability of Success in Primary Flight Training as a Function of ASTB Scores and API Grades: An Example of the Statistical Inferencing Component of the Pilot Prediction System.* NAMRL-1404, Naval Aerospace Medical Research Laboratory, Pensacola, FL, 1998.

2. Geisser, S. *An Introduction to Predictive Inference.* Chapman and Hall, New York, NY, 1993.

3. Press, S. J. *Bayesian statistics: principles, models, and applications.* John Wiley & Sons, New York, NY, 1989.

4. Berger, J. O. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, Berlin, Germany, 1985.

5. Coombs, C. H., Dawes, R. M., and Tversky, A. *Mathematical Psychology: An Elementary Introduction.* Prentice-Hall, Englewood Cliffs, NJ, 1970.

6. Blower, D. J. *Statistical Analysis of Performance Issues in Navy Flight Training.* NAMRL-1408, Naval Aerospace Medical Research Laboratory, Pensacola, FL, 2000.

7. Blower, D. J. *An Update to the LCAC Navigator Selection System Prediction Algorithm.* NAMRL Special Report 00-3, Naval Aerospace Medical Research Laboratory, Pensacola, FL, 2000.

8. Zellner, A. *An Introduction to Bayesian Inference in Econometrics.* John Wiley & Sons, New York, NY, 1971.

Intentionally blank

Appendix A

Formal Derivation of the Bayesian Predictive Density

Intentionally blank

This appendix contains a formal proof of the Bayesian predictive density function. The derivation starts out with a generic notation to make it easier to follow each individual step in the proof, and then switches at the end to the notation that we have used throughout the paper. The proof depends solely upon the product rule and the sum rule from probability theory and Bayes's theorem, which itself is, in turn, derived from the sum and product rules.

As a mental pump priming for the main proof, we first present a simpler proof which has all the necessary elements, but in a more digestible form. Consider the joint probability of two propositions A and B written as $P(A, B)$. Proposition B can be broken down into K mutually exclusive and exhaustive subpropositions $B_1, B_2, \cdots B_j \cdots B_K$. By the sum rule,

$$P(A) = \sum_{j=1}^{K} P(A, B_j). \tag{A1}$$

By the product rule, the joint probability in the rhs of Equation (A1) can be written as

$$P(A, B_j) = P(A|B_j)\ P(B_j) \tag{A2}$$

so we have from Equations (A1) and (A2),

$$P(A) = \sum_{j=1}^{K} P(A|B_j)\ P(B_j). \tag{A3}$$

Equation (A3) becomes more transparent when proposition A stands for the data D and subpropositions $B_1, B_2, \cdots B_j \cdots B_K$ for the K hypotheses. Then Equation (A3) reads

$$P(D) = \sum_{j=1}^{K} P(D|H_j)\ P(H_j) \tag{A4}$$

which is the denominator in Bayes's Theorem and explains the transition in the denominator from Equation (14) to Equation (15) in the main text.

The proof for the predictive density follows along the same lines, although it is a little trickier because of the many manipulations needed to get all the symbols in the right place. Instead of just two propositions A and B in the example above, we now have four propositions. I have found it easier to derive the predictive equation by first using a generic notation, w, x, y and z, for the four propositions, and then substituting the notation we have used throughout the paper at the final step.

To begin, write out the joint probability for all four propositions under consideration, $P(w, x, y, z)$, just as we did above. Again, following the same pattern as above, use the product rule to form

$$P(w, x, y, z) = P(w|x, y, z)\ P(x|y, z)\ P(y|z)\ P(z). \tag{A5}$$

Divide both sides by $P(z)$,

$$\frac{P(w, x, y, z)}{P(z)} = P(w|x, y, z)\ P(x|y, z)\ P(y|z). \tag{A6}$$

By Bayes's Theorem, the lhs of Equation (A6) is also equal to

$$\frac{P(w, x, y, z)}{P(z)} = P(w, x, y|z). \tag{A7}$$

Therefore, equating the rhs of Equations (A6) and (A7) yields

$$P(w, x, y|z) = P(w|x, y, z)\ P(x|y, z)\ P(y|z). \tag{A8}$$

Divide both sides of Equation (A8) by $P(y|z)$,

$$\frac{P(w,x,y|z)}{P(y|z)} = P(w|x,y,z)\ P(x|y,z). \tag{A9}$$

Now take the rhs of Equation (A9) and use the product rule in reverse,

$$P(w|x,y,z)\ P(x|y,z) = P(w,x|y,z). \tag{A10}$$

The lhs of Equation (A9) is equal to the rhs of Equation (A10). Therefore,

$$\frac{P(w,x,y|z)}{P(y|z)} = P(w,x|y,z). \tag{A11}$$

Use the sum rule on the rhs of Equation (A11)

$$\int P(w,x|y,z)\ dx = P(w|y,z). \tag{A12}$$

In the first derivation that began this appendix, the K discrete values of B were summed over in Equation (A1) to find A. This operation is known in the Bayesian jargon as "eliminating a nuisance parameter." In Equation (A12), the nuisance parameter x was eliminated. Because this parameter will turn out to be continuous, perform an integration instead of the summation. Substitute the lhs of Equation (A10) for the integrand in the lhs of Equation (A12),

$$P(w|y,z) = \int P(w|x,y,z)\ P(x|y,z)\ dx. \tag{A13}$$

For the final step, make w independent of y,

$$P(w|z) = \int P(w|x,z)\ P(x|y,z)\ dx. \tag{A14}$$

This completes the proof for the Bayesian predictive density where w is the new data and z is the state of the world for which this new data is assumed true. x is the parameter, or set of parameters, in the model which describe how the data are generated, and y is all the past data. Explicitly matching up this generic notation with the notation of the particular problem being treated in this paper, we have

$$\begin{aligned} w &\equiv D_{N+1} \\ x &\equiv \boldsymbol{\lambda} \\ y &\equiv D_N \\ z &\equiv \text{Pass or Fail} \end{aligned}$$

D_{N+1} stands for the data from the selection test for a particular candidate whose training outcome we wish to predict, D_N stands for the data from the selection test and known training outcomes for the validation sample consisting of the previous N subjects, and $\boldsymbol{\lambda}$ stands for the set of parameters in the model which tell us how the data came about. The conditioning information z can take on only two values, either PASS or FAIL. If $z = \text{PASS}$, then Equation (A14) looks like

$$P(D_{N+1}|\text{Pass}) = \int P(D_{N+1}|\boldsymbol{\lambda},\text{Pass})\ P(\boldsymbol{\lambda}|D_N,\text{Pass})\ d\boldsymbol{\lambda} \tag{A15}$$

in the standard notation used throughout the main text. Similarly, if $z = \text{FAIL}$, then Equation (A14) looks like

$$P(D_{N+1}|\text{Fail}) = \int P(D_{N+1}|\boldsymbol{\lambda}, \text{Fail})\; P(\boldsymbol{\lambda}|D_N, \text{Fail})\; d\boldsymbol{\lambda}. \tag{A16}$$

The likelihood ratio, $\mathcal{L}(x)$ demands that we form two predictive densities, one conditioned on the PASS group and one conditioned on the FAIL group.

This is what we set out to accomplish. In words, the final result of Equations (A15) or (A16) says to first set up the likelihood of the data from the candidate being tested as conditioned on the set of parameters, $P(D_{N+1}|\boldsymbol{\lambda})$, and then multiply by the posterior probability of the parameters as conditioned on all of the previous data, $P(\boldsymbol{\lambda}|D_N)$. This multiplication is done over the entire range that the parameters can assume, and then the results are summed. In other words, Equations (A15) and (A16) are asking for the average likelihood of D_{N+1} with the weighting function being the posterior probability of the parameters. In any case, the result is the Bayesian predictive density of the score, or scores, from the test battery for the candidate currently undergoing selection. Technically, Equations (A15) and (A16) should replace the curves shown in Figs. 1–3. Although, as shown in Appendix B, the rigorous application of Equations (A15) and (A16) for one common case results in curves that are very similar to the idealized Normal curves already studied.

If these predictive densities turn out to be too difficult to solve analytically, we can still obtain an answer through a numerical approximation to the integral. An example of a computer program to perform such a numerical approximation to the predictive density can be found in Blower [7]. The essence of the computer program is to calculate an average as just described. The likelihood of the composite score for the new candidate is calculated over many values of the parameters. If the number of parameters is small, as in the Normal case when there are only two parameters, $\boldsymbol{\lambda} = \{\mu, \sigma\}$, then a grid can be constructed over a reasonable range of the parameters. At each grid value, the posterior density function of the parameters is calculated. These values serve as the weights that form the average of the likelihood. After each likelihood is multiplied by its respective weight over many values of the parameters, this sum is divided by the sum of the weights. The resulting value is a good estimate of the integral in Equations (A15) and (A16).

After obtaining the predictive densities, the prediction algorithm can be written in its most general form. Because the likelihood ratio is equal to

$$\mathcal{L}(x) = \frac{P(D_{N+1}|\text{Pass})}{P(D_{N+1}|\text{Fail})},$$

Equation (21) in the main part of the paper can be reexpressed as

$$\frac{\int_R P(D_{N+1}|\boldsymbol{\lambda}, \text{Pass})\; P(\boldsymbol{\lambda}|D_N, \text{Pass})\; d\boldsymbol{\lambda}}{\int_R P(D_{N+1}|\boldsymbol{\lambda}, \text{Fail})\; P(\boldsymbol{\lambda}|D_N, \text{Fail})\; d\boldsymbol{\lambda}} \geq \frac{C_1}{C_2} \times \frac{P(\text{Fail})}{P(\text{Pass})}. \tag{A17}$$

Intentionally blank

Appendix B

An Analytical Solution for a Bayesian Predictive Density that Justifies the Use of the Normal Curve

Intentionally blank

This appendix shows how to derive an analytical solution for either one of the integrals on the lhs of Equation (A17). The solution is for the circumstances treated in the text as shown in Figs. 1–3 where we assumed Normal distributions for the predictive densities. The derivation of a predictive density like this is well known in the Bayesian literature. We follow closely the account given by Zellner [8].

Assume that both the data D_N from the previous N candidates and the data D_{N+1} provided by a new candidate are discriminant scores arising from the use of a discriminant analysis. Each candidate receives one discriminant score that is a linear weighting of the many test scores comprising the test battery. The weights are determined by the discriminant analysis. Such summary scores are usually called *composite scores* in the PPS. Discriminant analysis constructs these composite scores such that they are distributed according to a Normal pdf with a standard deviation of 1. The mean of the PASS group and the mean of the FAIL group are separated as much as possible given the data from the training outcomes.

If the composite score for the ith candidate is labeled as d_i, then the Normal pdf is

$$\text{Normal pdf} = \frac{1}{\sqrt{2\pi}\sigma_p} \exp\left\{ -\frac{1}{2} \left(\frac{d_i - \mu_p}{\sigma_p} \right)^2 \right\} \tag{B1}$$

just as in Equation (23). By construction, $\sigma_p = 1$, so the pdf can be shortened to

$$\text{Normal pdf} = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} (d_i - \mu_p)^2 \right\}. \tag{B2}$$

The validation sample consists of N_p candidates with composite scores $d_1, d_2 \cdots d_{N_p}$. These scores come from only those candidates who passed, as the subscript p indicates. Assuming independence of scores from the N_p different subjects,

$$\begin{aligned}
P(d_1, d_2 \cdots d_{N_p} | \mu_p, \sigma_p = 1) &= \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} (d_1 - \mu_p)^2 \right\} \times \\
&\quad \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} (d_2 - \mu_p)^2 \right\} \times \\
&\quad \vdots \\
&\quad \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} (d_{N_p} - \mu_p)^2 \right\} \\
&= \left(\frac{1}{\sqrt{2\pi}} \right)^{N_p} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N_p} (d_i - \mu_p)^2 \right\}
\end{aligned} \tag{B3}$$

The term

$$\left(\frac{1}{\sqrt{2\pi}} \right)^{N_p}$$

in Equation (B3) is a constant value for the fixed value of N_p and thus can be absorbed into a proportionality factor. Therefore, the likelihood is proportional to the exponential factor,

$$P(d_1, d_2 \cdots d_{N_p} | \mu_p, \sigma_p = 1) \propto \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N_p} (d_i - \mu_p)^2 \right\} \tag{B4}$$

We now show that

$$\sum_{i=1}^{N_p}(d_i - \mu_p)^2 = N_p(\mu_P - \overline{d})^2 + \text{Constant}$$

where $\overline{d}$ is the sample average composite score over the N_p candidates who passed.

$$\begin{aligned}
\sum_{i=1}^{N_p}(d_i - \mu_p)^2 &= \sum [\, (d_i - \overline{d}) - (\mu_p - \overline{d}) \,]^2 \\
&= \sum [\, (d_i - \overline{d})^2 - 2(d_i - \overline{d})(\mu_p - \overline{d}) + (\mu_p - \overline{d})^2 \,] \\
&= \sum (d_i - \overline{d})^2 - 2\sum (d_i - \overline{d})(\mu_p - \overline{d}) + \sum (\mu_p - \overline{d})^2 \\
&= \sum \, (d_i - \overline{d})^2 + \sum \, (\mu_p - \overline{d})^2 \\
&= \sum \, (d_i - \overline{d})^2 + N_p(\mu_p - \overline{d})^2 \\
&= N_p(\mu_p - \overline{d})^2 + \text{Constant.}
\end{aligned}$$

The first line subtracts a constant, the sample mean of the composite scores, $\overline{d}$, and then adds the same constant to the expression. Then in the second line, the expression in brackets is expanded by squaring. In the third line, the summation sign is distributed across the three terms in the expression. The middle term works out to be zero because

$$\begin{aligned}
\sum (d_i - \overline{d}) &= \sum d_i - N_p\overline{d} \\
&= N_p \times \left[\frac{\sum d_i}{N_p} - \frac{N_p\overline{d}}{N_p}\right] \\
&= N_p \times (\overline{d} - \overline{d}) \\
&= 0
\end{aligned}$$

which results in the fourth line. The first term is again a constant for any fixed set of data, being the sum of sample squared deviations. It can also be absorbed into the proportionality factor. So Equation (B4) can now be written as

$$P(d_1, d_2 \cdots d_{N_p} | \mu_p, \sigma_p = 1) \propto \exp\left\{-\frac{1}{2}N_p(\mu_p - \overline{d})^2\right\}. \tag{B5}$$

By Bayes's Theorem,

$$P(\mu_p, \sigma_p | D_N) \propto P(D_N | \mu_p, \sigma_p) \; P(\mu_p, \sigma_p)$$

Since $\sigma_p = 1$, the prior probability for the joint occurrence of the two parameters is usually assigned as a constant k in the Bayesian approach. The same range will be assigned for the joint occurrence of (μ_f, σ_f). Therefore, the posterior probability of the parameters is

$$\begin{aligned}
P(\mu_p, \sigma_p | D_N) &\propto P(D_N | \mu_p, \sigma_p) P(\mu_p, \sigma_p) \\
&\propto P(D_N | \mu_p, \sigma_p) \times k \\
&\propto \exp\left\{-\frac{1}{2}N_p(\mu_p - \overline{d})^2\right\}.
\end{aligned}$$

Solving the formula for the predictive density given the PASS group

$$\int_R P(D_{N+1}|\boldsymbol{\lambda}, \text{Pass}) P(\boldsymbol{\lambda}|D_N, \text{Pass})\, d\boldsymbol{\lambda}$$

is our eventual goal. We have just found one of those two terms, the posterior probability of the parameters, as

$$P(\boldsymbol{\lambda}|D_N, \text{Pass}) \propto \exp\left\{-\frac{1}{2}N_p(\mu_p - \overline{d})^2\right\}$$

where

$$\boldsymbol{\lambda} = \{\mu_p, \sigma_p\}.$$

From the same assumptions that started out this appendix, the likelihood of the new composite score, D_{N+1}, obtained from the candidate currently being tested, is

$$P(D_{N+1}|\mu_p, \sigma_p, \text{Pass}) \propto \exp\left\{-\frac{1}{2}(D_{N+1} - \mu_p)^2\right\}.$$

Therefore,

$$\begin{aligned}
\int_R P(D_{N+1}|\mu_p, \sigma_p)\, P(\mu_p, \sigma_p|D_N)\, d\mu\, d\sigma &= \int_R \exp\left\{-\frac{1}{2}(D_{N+1} - \mu_p)^2\right\} \times \exp\left\{-\frac{1}{2}N_p(\mu_p - \overline{d})^2\right\} d\mu\, d\sigma \\
&= \int_R \exp\left\{-\frac{1}{2}[(D_{N+1} - \mu_p)^2 + N_p(\mu_p - \overline{d})^2]\right\} d\mu\, d\sigma. \qquad \text{(B6)}
\end{aligned}$$

An analytical solution to this integral was worked out by Zellner [8]. In our notation, this solution is

$$P(D_{N+1}|\text{Pass}) = \frac{1}{\sqrt{2\pi(1 + 1/N_p)}} \exp\left\{-\frac{1}{2}\frac{(D_{N+1} - \overline{d})^2}{(1 + 1/N_p)}\right\}. \qquad \text{(B7)}$$

This is a Normal distribution for the composite score, D_{N+1}, as obtained from the new candidate we wish to classify correctly. It is centered on the sample mean of the composite scores, $\overline{d}$, with a variance inflated by a factor related to the sample size of the PASS group, $1 + \frac{1}{N_p}$. As you can see, when the sample sizes N_p and N_f become large, the predictive density functions are essentially the same as shown earlier in Figs 1–3. Those Normal curves were the same as Equation (B7) when $N \to \infty$. It is only when the sample sizes are small that Equation (B7) will provide the necessary correction for the likelihood ratio. Even in this case, the correction is much smaller than might have been expected.

In the earlier example, when the sample mean for the composite scores of the FAIL group occurred at -1.00 and the sample mean for the composite scores of the PASS group occurred at $+1.00$, the threshold score was placed at 0.00 for $\beta = 1$. Suppose that each sample mean was based on a small number of candidates. Let $N_p = 30$ and $N_f = 15$. Now we want to find that value of D_{N+1} such that

$$\mathcal{L}(x) = \frac{P(D_{N+1} =?|\text{Pass})}{P(D_{N+1} =?|\text{Fail})} = 1$$

or, equivalently, where

$$P(D_{N+1} =?|\text{Pass}) = P(D_{N+1} =?|\text{Fail}).$$

Using Equation (B7) and the analogous equation based on the FAIL group, we find that equality holds between the two predictive functions if $D_{N+1} = -.00022$. Therefore, there is only a very slight adjustment to the threshold score from its former value of 0 to accommodate the small sample sizes.

If β changes so that the threshold score is further out on the predictive density functions, then the change in the threshold score will be little larger for small sample sizes but still not very dramatic. For example, if $\beta = 3$, the threshold score is placed at $+0.55$ for the idealized case but only moves up to $+0.58$ when the correct formula taking account of the small sample sizes is used. Therefore, a new candidate would have to obtain a composite score $D_{N+1} \geq +0.58$ in order to be a predicted pass.

The moral of this appendix is that for composite scores based on any reasonable sample size, the predictive density for each group can be taken as the Normal distribution centered at the sample means of each group with a standard deviation of 1.00. The kind of thorough Bayesian analysis derived in this appendix justifies the standard practice illustrated in the final section of the main part of the paper.

If you are not willing to accept a standard deviation of 1.00 for the composite scores, Blower [7] has shown how the same Bayesian approach explained here can be used to take account of the uncertainty in both the sample means and sample standard deviations. As you might expect, the threshold score undergoes a greater adjustment in this case but still doesn't move very far from the idealized approximation presented earlier. In any case, one can compute a numerical approximation for the integrals in Equation (A17) for any situation where an analytical solution does not exist or is too hard to find. The likelihood ratio, $\mathcal{L}(x)$, can be found by forming the ratio of two such numerical approximations, and the prediction algorithm will continue to function under any circumstances.

Of course, since the algorithm is outputting a binary decision, only likelihood ratios very close to the threshold score would have to be calculated very exactly anyway. Or, stating it in another way more in line with the actual implementation, only composite scores very near to the threshold score would be affected in terms of a predicted pass or predicted fail.

Reviewed and approved 27 APR 01

R. R. STANNY, Ph.D.
Technical Director

This research was sponsored by the Office of Naval Research under work unit 62233N MM33P30-001-7801.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE 27 April 2001	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE The Statistical Foundations of the Pilot Prediction System	5. FUNDING NUMBERS 62233N MM33P30-001-7801
6. AUTHOR(S) D. J. Blower	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Aerospace Medical Research Laboratory 51 Hovey Road Pensacola, FL 32508-1046	8. PERFORMING ORGANIZATION REPORT NUMBER NAMRL 01-1
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Office of Naval Research 800 N. Quincy Street Arlington, VA 22242-5160	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE
Approved for public release; distribution is unlimited.	

13. ABSTRACT ***(Maximum 200 words)***

The Pilot Prediction System (PPS) is a research effort designed to provide Navy policy makers with improved access to selection and training data in the aviation community. One of its main features is the ability to make predictions about the future success of aviation candidates in flight training. The purpose of this report is to present in some detail the statistical foundations of this feature of the PPS. We first describe the rudiments of statistical decision theory. Such a theory allows us to make the very best decision possible when faced with the inherent uncertainty about what is actually going to take place in the future. The second pillar on which the PPS is built is the treatment of probability from the Bayesian perspective. Two technical appendices are included for the interested reader. The first contains a simplified proof of the Bayesian predictive density that allows the prediction algorithm in the PPS to be written in its most general form. The second shows an analytical solution derived from the theory developed in the first appendix justifying a practical approximation for predicting pass or fail during flight training.

14. SUBJECT TERMS Aviation Selection, Software Prototype for Predicting Aviation Success, Bayesian Statistics, Predictive Probability Density	15. NUMBER OF PAGES 35
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR